Inexact SARAH for Solving Stochastic Optimization Problems

Lam M. Nguyen, Katya Scheinberg, Martin Takáč

INFORMS Annual Meeting
November 6, 2018

IBM Research

LEHIGH UNIVERSITY

Problem Description

We consider the stochastic optimization problem:

$$\min_{w \in \mathbb{R}^d} \{F(w) = \mathbb{E}[f(w;\xi)]\}$$

Special case, finite-sum (with large *n*) problem:

$$\min_{w \in \mathbb{R}^d} \left\{F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)\right\}$$

Finite-sum Problem

Optimize a finite sum with large number of elements n

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \right\}$$

Training set: $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

f_i- **strongly convex**: linear regression, binary classification

ℓ_2-regularized least squares regression: $f_i(w) = \left(x_i^T w - y_i\right)^2 + \frac{\lambda}{2}\|w\|^2$

ℓ_2-regularized logistic regression: $f_i(w) = \log(1 + \exp(-y_i x_i^T w)) + \frac{\lambda}{2}\|w\|^2$

f_i- **nonconvex**: neural networks

Some "gradient" methods to solve this problem

"Full gradient": Gradient Descent

"Stochastic": SGD [H. Robbins & S. Monro, 1951]

"Variance Reduction": SAG [M. Schmidt et. al., 2013], SAGA [A. Defazio et. al., 2014], SVRG [R. Johnson and T. Zhang, 2013], SARAH [L. Nguyen et. al., 2017]

SARAH Algorithm

SARAH [Nguyen et. al., 2017]

- It also does **restarting** as SVRG [Johnson & Zhang, 2013]
- It takes **recursive** gradient estimator

Parameters: the learning rate $\eta > 0$ and the inner loop size m.
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \dots$ **do**
  $w_0 = \tilde{w}_{s-1}$
  $v_0 = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_0)$
  $w_1 = w_0 - \eta v_0$
  Iterate:
  for $t = 1, \dots, m-1$ **do**
    Sample i_t uniformly at random from $[n]$
    $v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}$
    $w_{t+1} = w_t - \eta v_t$
  end for
  Set $\tilde{w}_s = w_t$ with t chosen uniformly at random from $\{0, 1, \dots, m\}$
end for

SARAH Algorithm

SARAH [Nguyen et. al., 2017]

- It also does **restarting** as SVRG [Johnson & Zhang, 2013]
- It takes **recursive** gradient estimator

Parameters: the learning rate $\eta > 0$ and the inner loop size m.
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \ldots$ **do**
  $w_0 = \tilde{w}_{s-1}$
  $v_0 = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_0)$
  $w_1 = w_0 - \eta v_0$
  Iterate:
  for $t = 1, \ldots, m-1$ **do**
    Sample i_t uniformly at random from $[n]$
    $v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}$
    $w_{t+1} = w_t - \eta v_t$
  end for
  Set $\tilde{w}_s = w_t$ with t chosen uniformly at random from $\{0, 1, \ldots, m\}$
end for

Outer loop

SARAH Algorithm

SARAH [Nguyen et. al., 2017]

- It also does **restarting** as SVRG [Johnson & Zhang, 2013]
- It takes **recursive** gradient estimator

Parameters: the learning rate $\eta > 0$ and the inner loop size m.
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \ldots$ **do**
  $w_0 = \tilde{w}_{s-1}$
  $v_0 = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_0)$
  $w_1 = w_0 - \eta v_0$
  Iterate:
  for $t = 1, \ldots, m-1$ **do**
    Sample i_t uniformly at random from $[n]$
    $v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}$
    $w_{t+1} = w_t - \eta v_t$
  end for
  Set $\tilde{w}_s = w_t$ with t chosen uniformly at random from $\{0, 1, \ldots, m\}$
end for

Inner loop

Outer loop

SARAH Algorithm

SARAH [Nguyen et. al., 2017]

- It also does **restarting** as SVRG [Johnson & Zhang, 2013]
- It takes **recursive** gradient estimator

Parameters: the learning rate $\eta > 0$ and the inner loop size m.
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \ldots$ **do**

Full gradient computing

$w_0 = \tilde{w}_{s-1}$
$\boxed{v_0 = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_0)}$
$w_1 = w_0 - \eta v_0$
Iterate:
for $t = 1, \ldots, m-1$ **do**
Sample i_t uniformly at random from $[n]$
$v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}$
$w_{t+1} = w_t - \eta v_t$
end for

Inner loop

Set $\tilde{w}_s = w_t$ with t chosen uniformly at random from $\{0, 1, \ldots, m\}$
end for

Outer loop

SARAH Algorithm

SARAH [Nguyen et. al., 2017]

- It also does **restarting** as SVRG [Johnson & Zhang, 2013]
- It takes **recursive** gradient estimator

Parameters: the learning rate $\eta > 0$ and the inner loop size m.
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \dots$ **do**

$w_0 = \tilde{w}_{s-1}$

$\boxed{v_0 = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w_0)}$ Full gradient computing

$w_1 = w_0 - \eta v_0$

Iterate:
for $t = 1, \dots, m-1$ **do**

Sample i_t uniformly at random from $[n]$

$\boxed{v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}}$

$w_{t+1} = w_t - \eta v_t$

end for

Set $\tilde{w}_s = w_t$ with t chosen uniformly at random from $\{0, 1, \dots, m\}$

end for

Inner loop

Outer loop

SARAH update (stochastic gradient computing)

SARAH Algorithm

SARAH [Nguyen et. al., 2017]

- It also does **restarting** as SVRG [Johnson & Zhang, 2013]
- It takes **recursive** gradient estimator

Parameters: the learning rate $\eta > 0$ and the inner loop size m.
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \dots$ **do**
  $w_0 = \tilde{w}_{s-1}$
  $\boxed{v_0 = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_0)}$
  $w_1 = w_0 - \eta v_0$
  Iterate:
  for $t = 1, \dots, m-1$ **do**
    Sample i_t uniformly at random from $[n]$
    $\boxed{v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}}$
    $w_{t+1} = w_t - \eta v_t$
  end for
  Set $\tilde{w}_s = w_t$ with t chosen uniformly at random from $\{0, 1, \dots, m\}$
end for

Full gradient computing

Inner loop

Outer loop

SARAH update (stochastic gradient computing)

$$\boxed{v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_0) + v_0}$$

SVRG

SARAH One Outer Loop

Recall the update: $w_{t+1} = w_t - \eta v_t$

- P is L-smooth and μ-strongly convex

$$\mathbb{E}[||v_t||^2] \leq \rho^t \cdot \mathbb{E}[||\nabla F(w_0)||^2]$$

$$\rho = 1 - \left(\frac{2}{\eta L} - 1\right)\mu^2\eta^2 < 1, \qquad \eta < \frac{2}{L}$$

- Each $f_i, \forall i,$ is L-smooth and μ-strongly convex

$$\mathbb{E}[||v_t||^2] \leq \rho^t \cdot \mathbb{E}[||\nabla F(w_0)||^2]$$

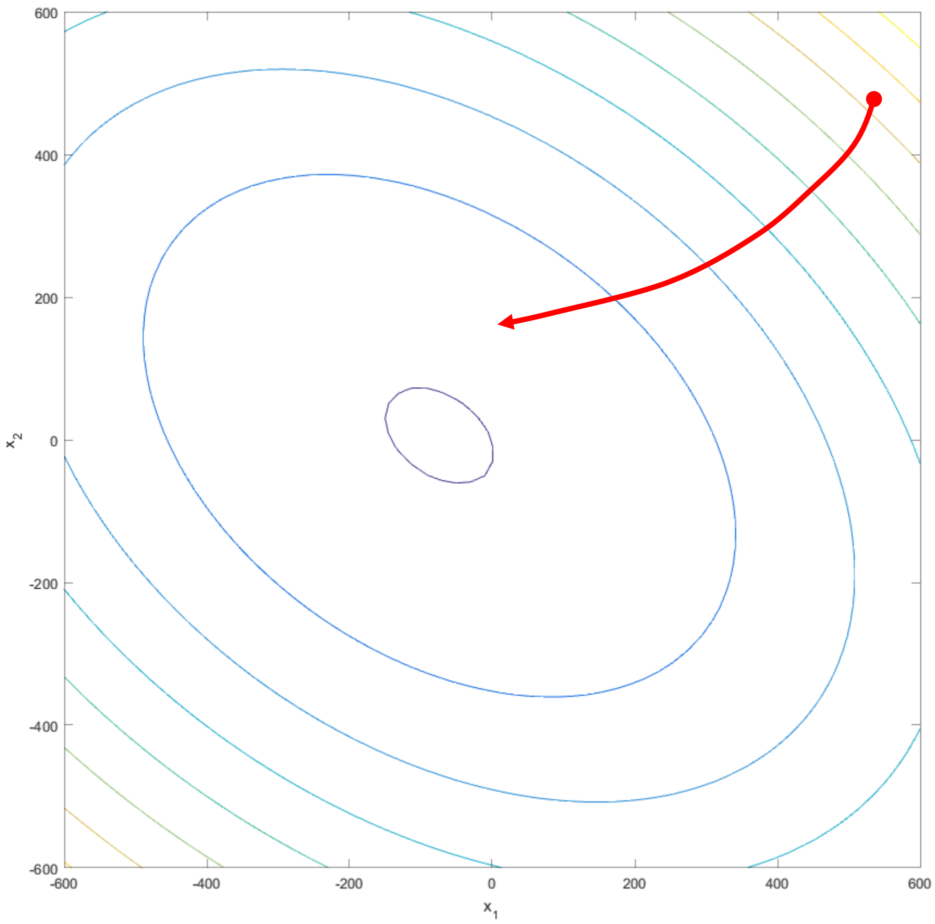
$$\rho = 1 - \frac{2\mu L\eta}{\mu + L} < 1, \qquad \eta \leq \frac{2}{L + \mu}$$

Hence,

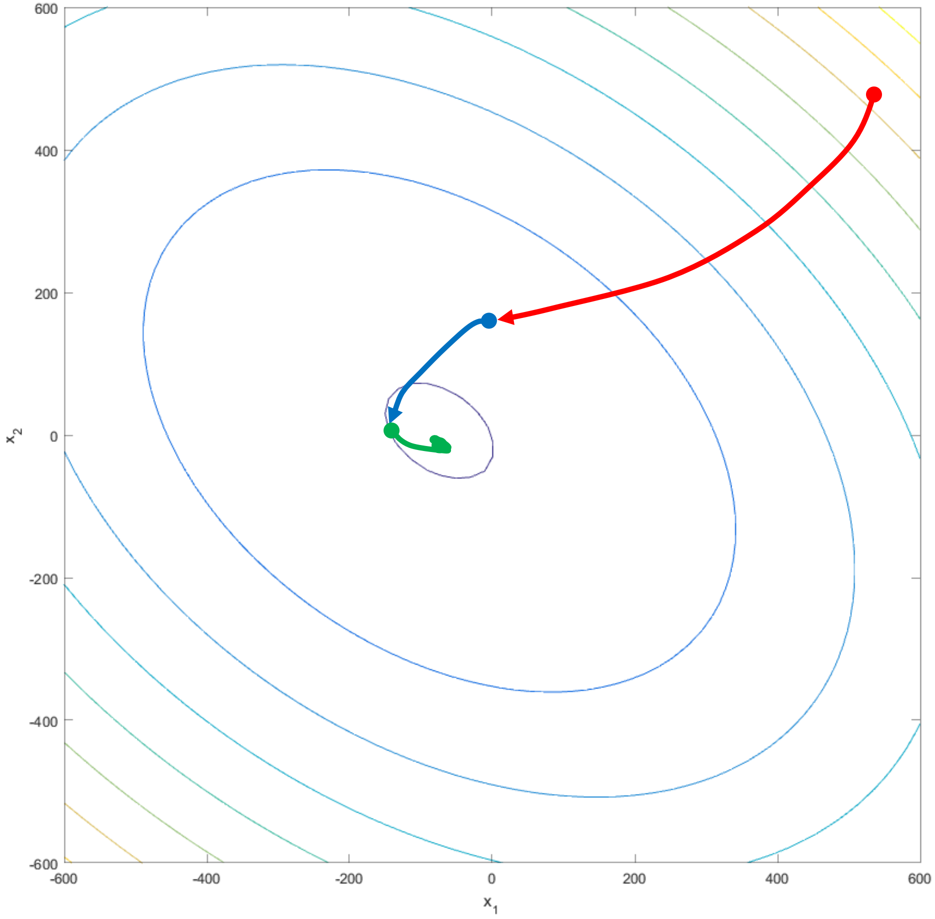
$$\mathbb{E}\left[||\boldsymbol{v}_t||^2\right] \to \mathbf{0} \Rightarrow \mathbb{E}\left[||\boldsymbol{w}_{t+1} - \boldsymbol{w}_t||^2\right] \to \mathbf{0}$$

SARAH is converging (somewhere) within a single outer loop with fixed “large” learning rate

SARAH Behavior



SARAH Behavior



Convergence Rates Comparisons

Strongly convex case: $\kappa = L/\mu$ is a condition number

Method	Complexity	Fixed Learning Rate	Low Storage Cost
GD	$\mathcal{O}\left(n\kappa \log\left(1/\epsilon\right)\right)$	✓	✓
SGD	$\mathcal{O}\left(1/\epsilon\right)$	✗	✓
SVRG	$\mathcal{O}\left(\left(n+\kappa\right)\log\left(1/\epsilon\right)\right)$	✓	✓
SAG/SAGA	$\mathcal{O}\left(\left(n+\kappa\right)\log\left(1/\epsilon\right)\right)$	✓	✗
SARAH	$\mathcal{O}\left(\left(n+\kappa\right)\log\left(1/\epsilon\right)\right)$	✓	✓

SGD: [Robbins & Monro, 1951], [Bottou et. al., 2018], [Nguyen et. al, 2018]
SVRG: [Johnson & Zhang, 2013]
SAG/SAGA: [Schmidt et. al., 2017], [Defazio et. al., 2014]
SARAH: [Nguyen et. al., 2017]

Problem Description

We consider the stochastic optimization problem:

$$\min_{w \in \mathbb{R}^d} \{F(w) = \mathbb{E}[f(w;\xi)]\}$$

Inexact SARAH (iSARAH)

Algorithm 1 Inexact SARAH (iSARAH)

Parameters: the learning rate $\eta > 0$ and the inner loop size m, the sample set size b.
Initialize: $\tilde{w}_0$.
Iterate:
for $s = 1, 2, \ldots, \mathcal{T}$, **do**
  $\tilde{w}_s = \text{iSARAH-IN}(\tilde{w}_{s-1}, \eta, m, b)$.
end for
Output: $\tilde{w}_{\mathcal{T}}$.

Algorithm 2 iSARAH-IN(w_0, η, m, b)

Input: $w_0 (= \tilde{w}_{s-1})$ the learning rate $\eta > 0$, the inner loop size m, the sample set size b.
Generate random variables $\{\zeta_i\}_{i=1}^{b}$ i.i.d.
Compute $v_0 = \frac{1}{b} \sum_{i=1}^{b} \nabla f(w_0; \zeta_i)$.
$w_1 = w_0 - \eta v_0$.
Iterate:
for $t = 1, \ldots, m-1$, **do**
  Generate a random variable ξ_t
  $v_t = \nabla f(w_t; \xi_t) - \nabla f(w_{t-1}; \xi_t) + v_{t-1}$.
  $w_{t+1} = w_t - \eta v_t$.
end for
Set $\tilde{w} = w_t$ with t chosen uniformly at random from $\{0, 1, \ldots, m\}$
Output: $\tilde{w}$

Inexact SARAH (iSARAH)

Algorithm 1 Inexact SARAH (iSARAH)

Parameters: the learning rate $\eta > 0$ and the inner loop size m, the sample set size b.
Initialize: $\tilde{w}_0$.
Iterate:
for $s = 1, 2, \ldots, \mathcal{T}$, **do**
  $\tilde{w}_s = \text{iSARAH-IN}(\tilde{w}_{s-1}, \eta, m, b)$.
end for
Output: $\tilde{w}_{\mathcal{T}}$.

Algorithm 2 iSARAH-IN(w_0, η, m, b)

Input: $w_0 (= \tilde{w}_{s-1})$ the learning rate $\eta > 0$, the inner loop size m, the sample set size b.
Generate random variables $\{\zeta_i\}_{i=1}^{b}$ i.i.d.
Compute $\boxed{v_0 = \frac{1}{b} \sum_{i=1}^{b} \nabla f(w_0; \zeta_i).}$ **NOT computing Full gradient**
$w_1 = w_0 - \eta v_0$.
Iterate:
for $t = 1, \ldots, m-1$, **do**
  Generate a random variable ξ_t
  $v_t = \nabla f(w_t; \xi_t) - \nabla f(w_{t-1}; \xi_t) + v_{t-1}$.
  $w_{t+1} = w_t - \eta v_t$.
end for
Set $\tilde{w} = w_t$ with t chosen uniformly at random from $\{0, 1, \ldots, m\}$
Output: $\tilde{w}$

Strongly Convex Results

Theorem 1: Suppose that $F(w)$ is μ-strongly convex and $f(w;\xi)$ is L-smooth and convex for every realization of ξ. Consider **Algorithm 1 (iSARAH)** with the choice of η, m , and b such that

$$\alpha = \frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2-\eta L} + \frac{4\kappa - 2}{b(2-\eta L)} < 1$$

(Note that $\kappa = L/\mu$). Then, we have

$$\mathbb{E}[||\nabla F(\widetilde{w}_s)||^2] - \Delta \leq \alpha^s(||\nabla F(\widetilde{w}_0)||^2 - \Delta)$$

where,

$$\Delta = \frac{\delta}{1-\alpha} \qquad \text{and} \qquad \delta = \frac{4}{b(2-\eta L)}\mathbb{E}[||\nabla f(w_*;\xi)||^2]$$

Strongly Convex Results

Theorem 1: Suppose that $F(w)$ is μ-strongly convex and $f(w;\xi)$ is L-smooth and convex for every realization of ξ. Consider **Algorithm 1 (iSARAH)** with the choice of η, m , and b such that

$$\alpha = \frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2-\eta L} + \frac{4\kappa - 2}{b(2-\eta L)} < 1$$

(Note that $\kappa = L/\mu$). Then, we have

$$\mathbb{E}[||\nabla F(\tilde{w}_s)||^2] - \Delta \leq \alpha^s(||\nabla F(\tilde{w}_0)||^2 - \Delta)$$

where,

$$\Delta = \frac{\delta}{1-\alpha} \qquad \text{and} \qquad \delta = \frac{4}{b(2-\eta L)}\mathbb{E}[||\nabla f(w_*;\xi)||^2]$$

Corollary 1: Let $\eta = \mathcal{O}\left(\frac{1}{L}\right), m = \mathcal{O}(\kappa), b = \mathcal{O}\left(\max\left\{\frac{1}{\epsilon}, \kappa\right\}\right), s = \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ in Theorem 1. Then, the total work complexity to achieve $\mathbb{E}[||\nabla F(\tilde{w}_s)||^2] \leq \epsilon$ is $\mathcal{O}\left(\left(\max\left\{\frac{1}{\epsilon}, \kappa\right\} + \kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$.

Nonconvex Results

Theorem 2: Suppose that $f(w;\xi)$ is L-smooth for every realization of ξ. Consider **Algorithm 2 (iSARAH-IN)** with

$$\eta \leq \frac{2}{L\left(\sqrt{1+4m}+1\right)} \leq \frac{1}{L} \qquad \text{and} \qquad b = \sqrt{m+1}$$

Then, we have

$$\mathbb{E}[||\nabla F(\widetilde{w}_s)||^2] \leq \frac{2}{\eta(m+1)}[F(w_0) - F^*] + \frac{1}{\sqrt{m+1}}\mathbb{E}[||\nabla f(w_0;\xi)||^2]$$

Nonconvex Results

Theorem 2: Suppose that $f(w;\xi)$ is L-smooth for every realization of ξ. Consider **Algorithm 2 (iSARAH-IN)** with

$$\eta \leq \frac{2}{L\left(\sqrt{1+4m}+1\right)} \leq \frac{1}{L} \qquad \text{and} \qquad b = \sqrt{m+1}$$

Then, we have

$$\mathbb{E}[||\nabla F(\widetilde{w}_s)||^2] \leq \frac{2}{\eta(m+1)}[F(w_0) - F^*] + \frac{1}{\sqrt{m+1}}\mathbb{E}[||\nabla f(w_0;\xi)||^2]$$

Corollary 2: Let $\eta = \mathcal{O}\left(\frac{1}{L}\right), b = \mathcal{O}\left(\sqrt{m+1}\right)$ in Theorem 2. Then, the total work complexity to achieve $\mathbb{E}[||\nabla F(\widetilde{w}_s)||^2] \leq \epsilon$ is $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$.

Convergence Rates Comparisons

For smooth strongly convex functions

Method	Bound	Problem type
SARAH (multiple loop)	$\mathcal{O}\left((n+\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$	Finite-sum
SVRG	$\mathcal{O}\left((n+\kappa)\log\left(\frac{1}{\epsilon}\right)\right)$	Finite-sum
SCSG	$\mathcal{O}\left(\left(\min\left\{\frac{\kappa}{\epsilon}, n\right\}+\kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$	Finite-sum
SCSG	$\mathcal{O}\left(\left(\frac{\kappa}{\epsilon}+\kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$	Expectation
SGD	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	Expectation
iSARAH (multiple loop)	$\mathcal{O}\left(\left(\max\left\{\frac{1}{\epsilon}, \kappa\right\}+\kappa\right)\log\left(\frac{1}{\epsilon}\right)\right)$	Expectation

For smooth nonconvex functions

Method	Bound	Problem type	Additional assumption
SARAH (one loop)	$\mathcal{O}\left(n+\frac{1}{\epsilon^2}\right)$	Finite-sum	None
SVRG	$\mathcal{O}\left(n+\frac{n^{2/3}}{\epsilon}\right)$	Finite-sum	None
SCSG	$\mathcal{O}\left(\min\left\{\frac{1}{\epsilon^{5/3}}, \frac{n^{2/3}}{\epsilon}\right\}\right)$	Finite-sum	Bounded variance
SCSG	$\mathcal{O}\left(\frac{1}{\epsilon^{5/3}}\right)$	Expectation	Bounded variance
SGD	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation	Bounded variance
iSARAH (one loop)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation	None

General Convex Results

Assumption: Let $\widetilde{w}_0, \widetilde{w}_1, \dots, \widetilde{w}_s$ be the outer iterations of **Algorithm 1 (iSARAH).** We assume that there exist $M > 0$ and $N > 0$ such that for all $k = 0,1, \dots, s$

$$F(\widetilde{w}_k) - F(w_*) \leq M||\nabla F(\widetilde{w}_k)||^2 + N$$

Theorem 3: $f(w; \xi)$ is L-smooth and convex for every realization of ξ. Consider **Algorithm 1 (iSARAH)** with the choice of η, m , and b such that

$$\alpha = \frac{2M}{\eta(m+1)} + \frac{\eta L}{2 - \eta L} + \frac{8LM - 1}{b(2 - \eta L)} < 1$$

(Note that $\kappa = L/\mu$). Then, we have

$$\mathbb{E}[||\nabla F(\widetilde{w}_s)||^2] - \Delta_c \leq \alpha^s(||\nabla F(\widetilde{w}_0)||^2 - \Delta_c)$$

where,

$$\Delta_c = \frac{\delta_c}{1 - \alpha_c} \quad \text{and} \quad \delta = \frac{2N}{\eta(m+1)} + \frac{8LN}{b(2 - \eta L)} + \frac{4}{b(2 - \eta L)} \mathbb{E}[||\nabla f(w_*; \xi)||^2]$$

Corollary 1: Let $\eta = \mathcal{O}\left(\frac{1}{L}\right), m = \mathcal{O}\left(\frac{1}{\epsilon}\right), b = \mathcal{O}\left(\frac{1}{\epsilon}\right), s = \mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ in Theorem 3. Then, the total work complexity to achieve $\mathbb{E}[||\nabla F(\widetilde{w}_s)||^2] \leq \epsilon$ is $\mathcal{O}\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$.

Convergence Rates Comparisons

For smooth general convex functions

Method	**Bound**	**Problem type**	**Additional assumption**
SCSG	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation	None
SGD	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation	Bounded variance
iSARAH (one loop)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	Expectation	None
iSARAH (multiple loop)	$\mathcal{O}\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$	Expectation	Assumption 4

References

- L. Bottou, F. E. Curtis, and J. Nocedal. *Optimization Methods for Large-scale Machine Learning*. SIAM Review, 2018
- A. Defazio, F. Bach, S. Lacoste-Julien. *SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives*. NIPS 2014
- R. Johnson and T. Zhang. *Accelerating Stochastic Gradient Descent using Predictive Variance Reduction*. NIPS 2013.
- N. Le Roux, M. Schmidt, and F. Bach. *A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets*. NIPS 2012
- L. Lei and M. Jordan. *Less than a Single Pass: Stochastically Controlled Stochastic Gradient*. AISTATS 2017
- L. Lei, C. Ju, J. Chen, and M. I. Jordan. *Non-convex finite-sum optimization via SCSG methods*. NIPS 2017
- H. Robbins and S. Monro. *A Stochastic Approximation Method*. 1951
- L. Nguyen, J. Liu, K. Scheinberg, and M. Takac. *SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient*. ICML 2017
- L. Nguyen, P. H. Nguyen, M. van Dijk, P. Richtarik, K. Scheinberg, and M. Takac. *SGD and Hogwild! Convergence Without the Bounded Gradients Assumption*. ICML 2018
- S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. J. Smola. *Stochastic variance reduction for nonconvex optimization*. ICML 2016
- M. Schmidt, N. Le Roux, and F. Bach. *Minimizing finite sums with the stochastic average gradient*. Mathematical Programming 2017

SARAH

THANK YOU !!!

Lam M. Nguyen

LamNguyen.MLTD@gmail.com

https://lamnguyen-mltd.github.io/